THE FEDERAL COURTS LAW REVIEW

Volume 7, Issue 1 2014

Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review"

Maura R. Grossman[†] and
Gordon V. Cormack[††]

ABSTRACT

Approaches to technology-assisted review ("TAR") and its validation—presented as "obligations" under Federal Rule 26(g) in a recent article by Karl Schieneman and Thomas C. Gricks III—could, if prescribed, impair the effectiveness of TAR and increase its expense, thereby compromising a primary objective of Rule 26(g): to ensure a reasonable production at a proportionate cost.

Extraordinary culling efforts to enrich the collection are likely to eliminate large amounts of responsive information, while affording only the illusion of improved review or validation effectiveness. In addition, empirical evidence shows that the use of random selection for seed or training sets is inferior to keyword selection, and substantially inferior to the use of active learning—non-random selection determined by a machine-learning algorithm. Finally, exclusive focus on a particular statistical test, applied to a single phase of a review effort, does not provide adequate assurance of a reasonable production, and may be unduly burdensome. Validation should consider all available evidence concerning the effectiveness of the end-to-end review process, including prior scientific evaluation of the TAR method, its proper application by qualified individuals, and proportionate *post hoc* sampling for confirmation purposes.

TABLE OF CONTENTS

† A.B., Brown University; M.A. and Ph.D. in Psychology, Gordon F. Derner Institute of Advanced Psychological Studies, Adelphi University; J.D., Georgetown University Law Center; Of Counsel, Wachtell, Lipton, Rosen & Katz. The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

†† B.Sc., M.Sc., and Ph.D. in Computer Science, University of Manitoba; Professor and Co-Director of the Information Retrieval Group, David R. Cheriton School of Computer Science, University of Waterloo.

I. INTRODUCTION

In a recent Federal Courts Law Review article titled "The Implications of Rule 26(g) on the Use of Technology-Assisted Review,"[1] Karl Schieneman and Thomas C. Gricks III address important issues surrounding the use of technology-assisted review ("TAR") in electronic discovery. Schieneman and Gricks assert that Federal Rule of Civil Procedure 26(g) ("Rule 26(g)") imposes "unique obligations"[2] on responding parties that use TAR to produce documents in response to discovery requests. These obligations include:

1. Exercising "greater care in the collection of ESI" to "optimize and manage the richness of the database," by narrowing or culling the document collection "in order to [maximize] the fraction of relevant documents processed into the tool";[3] and

2. Training the TAR system using a random "seed" or "training" set, as opposed to one relying on judgmental sampling, which "may not be representative of the entire population of electronic documents within a given collection."[4]

Schieneman and Gricks' assertion appears to be driven by their premise that the "goal of technology-assisted review is to achieve an acceptable level of recall and precision, in light of proportionality considerations, based on a statistical quality control analysis."[5] They

---

1. Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 239 (2013), *available at* http://www.fclr.org/fclr/articles/html/2010/Gricks.pdf (hereinafter "Schieneman & Gricks").

2. *Id.* at 240.

3. *Id.* at 249, 251–52. "ESI" refers to "Electronically Stored Information." Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review with Foreword by John M. Facciola, U.S. Magistrate Judge*, 7 Fed. Cts. L. Rev. 1 (2013), at 15, 16, *available at* http://www.fclr.org/fclr/articles/html/2010/grossman.pdf (hereinafter "*TAR Glossary*").

4. Schieneman & Gricks, *supra* note 1, at 260.

5. *Id.* at 269. "Recall" is defined as "The fraction of Relevant Documents that are identified as Relevant by a search or review effort." *TAR Glossary*, *supra* note 3, at 27. "Precision" is defined

state that, to demonstrate "reasonable inquiry" under Rule 26(g), it is necessary to conduct random sampling sufficient to estimate recall and precision with a margin of error of at most ±5%, and a confidence level of at least 95%.[6]

We submit that the goal of technology-assisted review, or any other review,[7] is to identify for production as much responsive information as reasonably possible, at a proportionate cost. While statistics such as recall and precision are *measures* of success in achieving that goal, they are not the goal in itself. Indeed, when statistical outcomes become goals, they are subject to Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."[8] Schieneman and Gricks' asserted obligations would have the statistical tail wagging the best-practices dog; that is, they would have the TAR process dictated by sampling practices rather than by what would achieve the best possible review.

We are concerned that Schieneman and Gricks' prescriptions may offer false assurance that by rote adherence to a particular formulaic approach, a responding party can meet the "reasonable inquiry" requirement of Rule 26(g). Adoption of their prescriptions could also invite a requesting party to demand evidence of adherence "at every step of the process,"[9] leading to discovery about discovery and successive motions practice. Moreover, in many circumstances, rigid adherence to Schieneman and Gricks' proposals would incur disproportionate burden and cost, while compromising the effectiveness of—or even precluding—perfectly reasonable approaches to TAR and to validation.

Schieneman and Gricks appear to presuppose a particular approach to TAR, which we refer to here as "Simple Passive Learning" or "SPL," and further, to conflate this approach with a particular approach to validation. But there is no single approach to TAR or to validation, nor any necessity that TAR and validation employ a common approach.

We use, as an example of a TAR method that does not fit Schieneman and Gricks' assumptions, the method employed by Gordon V. Cormack and Mona Mojdeh in the course of their participation in the

---

(footnote continued)

as "The fraction of Documents identified as Relevant by a search or review effort that are in fact Relevant." *Id.* at 25.

6. Schieneman & Gricks, *supra* note 1, at 270 ("Regardless of the specific measure being evaluated, utilizing a sample having a confidence level of either 95% or 99%, and a nominal confidence interval of between ±2% and ±5% will satisfy the reasonable inquiry requirements of Rule 26(g).").

7. While we focus our comments on TAR, we dispute the logic by which Schieneman and Gricks conclude that the purpose, conduct, or validation of TAR methods differs from that of traditional review methods so as to incur "unique obligations."

8. Marilyn Strathern, *'Improving Ratings': Audit in the British University System*, 5 EUR. REV. 305, 308 (1997), *available at* http://journals.cambridge.org/article_S1062798700002660.

9. Schieneman & Gricks, *supra* note 1, at 248 ("[A]dherence to the principles of Rule 26(g) is critical at every step of the process when technology-assisted review is used. . . ." *See also id.* at 247 ("Ultimately, whether counsel has adequately discharged applicable Rule 26(g) obligations will typically be 'a fact intensive inquiry that requires evaluation of the procedures the producing party adopted during discovery. . . .'" (footnote omitted)).

TREC 2009 Legal Track,[10] and subsequently employed by the authors in fifty civil and regulatory matters. We refer to this method as "Continuous Active Learning" or "CAL."

We argue that the use of TAR—whether employing SPL or CAL—does not occasion the use of extraordinary efforts to "optimize and manage [] richness," and that such efforts may compromise not only the quality and cost effectiveness of the review, but also the accuracy and cost effectiveness of validation. We also argue that the use of TAR does not require the use of random seed or training sets, and that the removal of judgmental and other non-random input may impair the quality and increase the cost of the review. Finally, we argue that the proposed validation method—setting a recall target during "stabilization" and sampling the "review set" to ensure that the target is met[11]—incorrectly focuses on an intermediate phase of the end-to-end review process—a phase that may be a necessary part of SPL, but is absent from CAL. Validation, we argue, is best achieved by considering the end-to-end effectiveness of the review, and evaluating the totality of the evidence derived from multiple sources, not by considering only a single target measure applied to a particular phase of the review process.

We illustrate our arguments by considering the application of alternative TAR and validation approaches within the context of a hypothetical matter:

> The responding party has employed customary practices to identify custodians and ESI sources that may contain responsive information. After de-NISTing,[12] deduplication, and date restriction, one million documents that meet the criteria have been imported into a review tool. Unbeknownst to the responding party, the collection contains ten thousand responsive documents. The goal of the review is to find as many of these responsive documents as possible, at a proportionate cost, while ensuring through reasonable inquiry that this goal has been met.

---

10. Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS (2009), at 2–6, *available at* http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf (hereinafter "Cormack & Mojdeh"). *See also* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), at 31–34, *available at* http://jolt.richmond.edu/v17i3/article11.pdf.

11. *See* Schieneman & Gricks, *supra* note 1, at 263–73.

12. "De-NIST[ing]" is defined as "The use of an automated filter program that screens files against the NIST list in order to remove files that are generally accepted to be system generated and have no substantive value in most instances." The Sedona Conference, *The Sedona Conference Glossary: E-Discovery & Digital Information Management* (4th ed. 2014), at 13, *available at* https://thesedonaconference.org/publication/The Sedona Conference%C2%AE Glossary.

The remainder of this paper is organized as follows. Section II presents CAL and the evidence that it works at least as well as any TAR method reported in the scientific literature. Section III describes SPL and contrasts it with CAL. Section IV argues that extraordinary efforts to "optimize and manage the richness of the database" provide only the illusion of improved recall and easier validation; an apples-to-apples comparison reveals that such efforts are likely to reduce the number of responsive documents found by the review, without any commensurate reduction of "uncertainty in the result."[13] Section V presents evidence that modifying CAL to use random training examples actually harms its effectiveness, and that the use of judgmental training examples can improve SPL. Section VI discusses the need to validate the end-to-end effectiveness of the entire review effort, not just the specific portion of the review process that Schieneman and Gricks refer to as "technology-assisted review."[14] Section VII discusses the limitations of recall as a validation measure, the statistical and methodological challenges of measuring recall, and the mathematical unsoundness of the various sampling methods advanced by Schieneman and Gricks, as well the closely related "eRecall" method.[15] In Section VIII, we offer some suggestions for a way forward. Relying on the elements of the *Daubert* test, we argue that the principal focus of validation should be on (i) prior scientific evaluation of the TAR method, (ii) ensuring its proper application by qualified individuals, and (iii) proportionate *post hoc* sampling for confirmation purposes. Section IX offers our conclusions.

II. CONTINUOUS ACTIVE LEARNING

The TAR method found by the authors to be most effective is Continuous Active Learning ("CAL"). CAL involves two interactive tools: a keyword search system with relevance ranking,[16] and a machine-learning algorithm.[17] At the outset of the TAR process, keyword searches are performed and some of the top-ranked documents from each search are coded by a human reviewer as responsive or not. These coded documents (the "seed set") are used to train the learning algorithm, which ranks each document in the collection by the likelihood that it

13. Schieneman & Gricks, *supra* note 1, at 251–52.

14. *Id.* at 269.

15. Herbert L. Roitblat, *A Tutorial on Sampling in Predictive Coding* (OrcaTec LLC 2013), at 3, *available at* http://orcatec.com/2013/10/07/a-tutorial-on-sampling-in-predictive-coding/ ("[W]e call Recall estimated from Elusion and Prevalence eRecall to distinguish it from Recall computed directly. . . ."); Herbert L. Roitblat, *Measurement in eDiscovery: A Technical White Paper* (OrcaTec LLC 2013), at 10, *available at* http://orcatec.com/2013/10/06/measurement-in-ediscovery-a-technical-white-paper/ (hereinafter "*Measurement in eDiscovery*") ("Estimating Recall from Elusion can be called eRecall.").

16. "Relevance Ranking" is defined as "A search method in which the results are ranked from the most likely to the least likely to be Relevant to an Information Need; the result of such ranking. Google Web Search is an example of Relevance Ranking." *TAR Glossary*, *supra* note 3, at 28.

17. "Machine Learning" is defined as "The use of a computer Algorithm to organize or Classify Documents by analyzing their Features." *Id.* at 22.

contains responsive information. The top-ranked documents that have not yet been coded are then coded by a human reviewer, and the learning algorithm is retrained using all coded documents. This process of training and coding is repeated until the number of top-ranked documents containing responsive information drops precipitously. At that time, quality-assurance efforts, including, but not limited to, additional keyword searches, score modeling,[18] and random sampling, are undertaken. If any of these efforts uncovers more documents containing responsive information, the CAL process is restarted, provided that the quantity, novelty, and importance of the newly identified information justify the additional effort.

CAL was shown, in what Schieneman and Gricks refer to as the "preeminent study of the effectiveness of information retrieval techniques in the legal field"[19] (the "JOLT Study"[20]), to be superior to manual review for four review tasks conducted at TREC 2009.[21] Yet CAL would fail Schieneman and Gricks' Rule 26(g) test. In the four tasks reported in the JOLT Study, Cormack and Mojdeh achieved superior results applying CAL directly, without "optimiz[ing] and manag[ing] the richness" of the collections, even though each collection had "low richness," as defined by Schieneman and Gricks[22] (*i.e.*, 0.3%, 0.7%, 0.3%, and 1.5%).[23] Of the matters in which the authors have subsequently applied CAL—without "optimiz[ing] and manag[ing] [] richness"—more than 90% have involved collections with "low richness."[24] Moreover, at TREC 2009, and in all of the authors' subsequent matters, keyword search (*i.e.*, "judgmental selection"[25]) was used to create the seed set.

---

18. *See* Cormack & Mojdeh, *supra* note 10, at 7 (Figure 4 and accompanying text).

19. Schieneman & Gricks, *supra* note 1, at 263 ("The starting point for the analysis of technology-assisted review is the consideration of the relative effectiveness of available alternatives. The preeminent study of the effectiveness of information retrieval techniques in the legal field was the JOLT Study, published in 2011.").

20. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), *available at* http://jolt.richmond.edu/v17i3/article11.pdf.

21. Bruce Hedin, Stephen Tomlinson, Jason R. Baron & Douglas W. Oard, *Overview of the TREC 2009 Legal Track*, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS, at 17 Table 6 (Topics 201, 202, 203, and 207), *available at* http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf (hereinafter, "Hedin et al."). It is worth noting that e-discovery service provider H5, using a radically different, rule-based method—which also does not appear to conform to Schieneman and Gricks' asserted Rule 26(g) obligations—did equally well on a fifth review task at TREC 2009. *Id.* (Topic 204).

22. Schieneman & Gricks, *supra* note 1, at 250 ("a prevalence of only 1% (a low prevalence) . . . .").

23. Hedin et al., *supra* note 21, at 16 Table 5 (Topics 201, 202, 203, and 207).

24. Our median richness has been less than 1%; we have had very few matters with richness above 3%. *Compare with* Schieneman & Gricks, *supra* note 1, at 249 ("Richness tends to be between five percent (5%) and ten percent (10%) of the total collection, but may be greater or even an order of magnitude less," citing *Measurement in eDiscovery*, *supra* note 15, at 6 ("We tend to see that around 5 [sic] 5-10% of the documents in an average collection are responsive, though we have seen higher (up to around 50%) and lower (around 0.5%) Prevalence on occasion. Others report even lower Prevalence.")).

25. Schieneman & Gricks, *supra* note 1, at 259.

The hypothetical matter represents a situation that would be typical for the authors' application of TAR. The richness of the collection is 1% (10,000 responsive documents in a collection of 1,000,000 documents). We would expect to find representative exemplars within several hours of searching, and then employ an active-learning algorithm until most of the responsive documents were found. On the order of 20,000 documents would need to be reviewed before the bulk of the responsive documents were identified.[26] We would know that our approach was reasonably likely to succeed, because we had employed the same technique fifty times to a variety of collections and production requests, always achieving a successful result. During the conduct of the review, internal measures would reinforce our expectations, and after the review, sampling would further corroborate our belief that a reasonable review had been conducted.

III. SIMPLE PASSIVE LEARNING

The TAR method that appears to be assumed by Schieneman and Gricks, and many others, is Simple Passive Learning ("SPL"). A review involving SPL has two distinct phases: training and review. In the training phase, a set of training documents (often referred to as the "seed set") is identified and coded as responsive or not. At the end of the training phase, the coded training documents are used as input to a learning algorithm, which either identifies a subset of the collection as likely responsive (the "review set"), or scores each document in the collection by the likelihood—as estimated by the learning algorithm—that it is responsive, in which case the review set consists of the documents achieving a score equal to or exceeding some "threshold" or "cutoff" value. In the review phase, the review set is coded manually, and only those documents coded as responsive (less any coded as privileged) are produced. On rare occasions, the review phase may be omitted, and the review set produced without further review.

SPL methods diverge from one another in how the training documents are selected. The principal approaches are judgmental selection (*e.g.*, keyword search), random selection, or a combination of both. Judgmental selection relies on the skill of the searcher to find appropriate training examples; random selection relies on chance; a combination hedges its bets.

SPL differs from CAL in two important respects: In SPL, the learning algorithm plays no role in the selection of training examples, and the review phase is separate and distinct from the training phase. As a consequence, the size and composition of the training set play a critical role in SPL; once these choices are made and the training set is coded, the review set is fixed, and the quality of the review is largely

26. In our experience, it is generally necessary to review about twice as many documents as the number of responsive documents in the collection.

predetermined. In SPL, the critical step of determining whether a particular training set (and hence review set) is adequate, or whether it could be enhanced sufficiently to justify additional training effort, is known as "stabilization."[27]

To achieve stabilization, SPL methods typically employ an iterative training phase in which a candidate training set is constructed (using one of the possible selection methods) and used by the learning algorithm to create a candidate review set. If sampling indicates that the review set is "adequate," the training phase is complete; otherwise, the training set is revised (typically by adding more examples), a new candidate review set is created, and sampling is repeated until stabilization is achieved.

A precise definition of an "adequate review set" remains elusive, as does the design of a sampling strategy to determine adequacy, however defined. As suggested by Schieneman and Gricks, one might specify target levels of precision and recall, informed by what could reasonably be achieved with proportionate effort.[28] Unfortunately, determining what could reasonably be achieved involves predicting the future: How much could the review set be improved, for a given amount of additional training effort? If the targets are set too high, stabilization might never occur; if the targets are set too low, stabilization might occur too early, resulting in an inferior review compared to what could have been achieved with reasonable additional effort.

In our hypothetical matter, target levels of 75% might be set for both precision and recall. We might begin with a seed set of 1,000 documents identified using random selection, keyword search, or a combination of both. These documents would be used to train the learning algorithm and to identify a candidate review set. A random sample of 1,000 documents might then be drawn from the collection and used to estimate the precision and recall of the review set. Assuming the estimated precision or recall were inadequate (*i.e.*, < 75%), the sample would be added to the training set, and the process repeated, until both the estimated precision and recall exceeded 75%. At this point, the

---

27. *Id.* at 263. Schieneman and Gricks do not define the term "stabilization" in their article. Stabilization is a vendor term that has been used to refer to the point at which further training will not improve the effectiveness of the learning algorithm. *See, e.g.*, Chris Dale, *Far From the Black Box: Explaining Equivio Relevance to Lawyers – White Paper* (Equivio 2012), at 9, *available at* http://www.equivio.com/files/files/White_Paper_–_Far_from_the_Black_Box:_Explaining_Equivio Relevance_to_Lawyers.pdf ("[T]here comes a point when further training adds nothing to the system's understanding. This is known in Equivio Relevance as the 'stabilisation point'. . . ."). *See also Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, 187 (S.D.N.Y. 2012) (discussing stabilization in the context of "the training of the [TAR] software").

28. Schieneman & Gricks, *supra* note 1, at 266 ("Ultimately, what constitutes an acceptable level of recall will be based on the totality of the circumstances and, absent agreement, counsel should be prepared to make a proper showing under Federal Rule 26(b)(2)(C) to support any proposed target or actual result."). *See also id.* at 263 ("Perhaps the most critical question attendant to the use of technology-assisted review for the production of documents is this: what levels of recall and precision are sufficient under Federal Rule of Civil Procedure 26(g)? Not surprisingly, given the incipience of technology-assisted review as a document review and production technique, neither the case law nor the Federal Rules provide a bright-line answer."), 267 ("[T]he party proposing the limitation must demonstrate the need to limit that discovery (*i.e.*, establishing a recall objective)." (footnote omitted)).

training phase would be deemed complete, and the review phase would begin, using the final candidate review set.

Although not envisioned in the TAR process described by Schieneman and Gricks, sampling following the review phase could be used to confirm the expectation that a reasonable review had been conducted. *See* Section VI.

IV. COLLECTION ENRICHMENT

As illustrated in this section, an obligation to enrich the collection would increase the complexity and cost of the review, while providing only the illusion of improved validation. At the start of the review, it would be necessary for the responding party to sample the collection to estimate richness, so as to determine how much, if any, enrichment was required. In order to enrich the collection, it would be necessary to discard the vast majority of the documents, at least some of which would almost certainly be responsive and, by virtue of being discarded prior to the TAR effort, never reviewed or produced. To verify that a negligible number of responsive documents were discarded through enrichment, it would be necessary to sample the discarded documents—an even more challenging task than sampling the original collection, because the discard pile would have substantially lower richness. Moreover, target measures derived from the enriched collection would not be the same as—indeed, would be incomparable to—target measures derived from the original collection. As a consequence, it is entirely possible, as illustrated below, that a review effort achieving a higher recall on the enriched collection might actually find fewer responsive documents overall—and incur a higher level of uncertainty as to the quality of the result—than a review effort achieving a lower recall on the original collection. In short, collection enrichment simply moves the statistical goalposts.

In our hypothetical matter, a random sample would be used to estimate the richness of the original collection in order to determine how much enrichment would be necessary. Determining the sample size for this estimate would not be a trivial matter, as the choice would require a guess as to the richness, which, of course, is not known prior to taking the sample. The choice would further require an assessment of the responding party's tolerance for error in the estimate. An inadequate sample size could result in the estimate being either too low or too high: Too low an estimate would mislead the responding party to undertake excessive enrichment, while too high an estimate would mislead the responding party to undertake insufficient enrichment. A full discussion of sampling for this purpose is beyond the scope of this paper; for the purpose of illustration, we will arbitrarily choose a sample size of N=1,000. For this value of N, given the fact that (unbeknownst to the responding party) 10,000 of the 1,000,000 documents are responsive, approximately 10—but probably not exactly 10—of the documents in the sample would be responsive. For simplicity, let us assume that the

responding party were lucky and the number were exactly 10. The point estimate for richness would therefore be 10/1,000, or 1%.

In order to increase richness to 10%—a tenfold increase—it would be necessary to reduce the size of the collection at least tenfold; that is, to discard at least nine-tenths of the collection, or 900,000 documents. Assuming, for simplicity, that all of the discarded documents were truly non-responsive (*i.e.*, that *none* of the 10,000 responsive documents were among those discarded), the enriched collection would then have 100,000 documents, of which 10,000 were responsive (*i.e.*, 10% richness). This assumption, however, is not realistic; in reality, the discard pile would contain at least *some* percentage of responsive documents. Suppose that percentage were only one-third of 1% (0.33%). In that case, the enrichment process would discard 3,000 responsive documents and would yield an enriched collection with 100,000 documents, of which 7,000 were responsive (*i.e.*, 7% richness). By the same token, if, instead, two-thirds of 1% (0.67%) of the discard pile were responsive, 6,000 responsive documents would be lost, and the resulting enriched collection would have 4,000 responsive documents (*i.e.*, 4% richness). Clearly, it is important to know how many of the discarded documents are responsive, not only to achieve the desired richness, but, more importantly, to ensure that the baby has not been thrown out with the bathwater. Estimating the number of discarded responsive documents is fraught with exactly the same statistical peril—low richness—that optimizing and managing richness is intended to avoid.

Suppose, in our hypothetical matter, the richness of the enriched collection was 7% (*i.e.*, the enriched collection contained 7,000 responsive documents). Suppose further that 7% richness, although falling short of the 10% target, was deemed to be an adequate richness to proceed with TAR. If a TAR method were then applied to the enriched collection, achieving 80% recall on that collection, the result would be that 80% of the 7,000 responsive documents (or 5,600 responsive documents) would be identified by the enrichment and review processes combined. In contrast, if the same TAR method were applied directly to the original collection, achieving only 70% recall—instead of 80% recall—the result would be to find 70% of 10,000 documents, or 7,000 responsive documents. In short, the enrichment process offers the illusion of better recall (80% as compared to 70%), while actually identifying fewer responsive documents (5,600 as compared to 7,000). The comparison of the two recall values is specious, as they are derived from different collections. A valid comparison demands that both recall values be derived from the original collection, in which case we see that, in this simple example, the recall of TAR alone is 70%, while the recall of enrichment plus TAR is 56%.

We posit that the impetus for Schieneman and Gricks' advocacy of collection enrichment is the presumption that all TAR methods use SPL and, in particular, require precise statistical estimates to achieve

stabilization.[29] Only in light of this presumption can we make sense of Schieneman and Gricks' assertions that (i) "technology-assisted review depends primarily upon statistics to validate effectiveness";[30] (ii) "Rule 26(g) requires counsel to recognize, understand, and endeavor to avoid high levels of uncertainty, and minimize the margin of error, in their use of technology-assisted review";[31] and (iii) "these obligations require counsel to understand the impact of richness on the ability of technology-assisted review to cull relevant documents from the database, and to develop a strategy for avoiding as much uncertainty in the result as is reasonably possible."[32] As noted in Section II, CAL consistently achieves high recall from low-prevalence collections, and *does not* in any way depend on using statistics yielding a low margin of error. In contrast, as outlined in Section III, SPL *does* use statistics to determine when stabilization has occurred, so as to terminate the training phase. A high level of uncertainty could, in this situation, result in a premature end to training, an inadequate review set, and ultimately, an inadequate production. If an SPL method—or any TAR method—does not work well for collections with low prevalence, it should not be applied to collections with low prevalence. Extraordinary culling effort for the purpose of enrichment, in ways "not previously associated with traditional review techniques,"[33] is not the answer.

V. RANDOM TRAINING

It is difficult to determine precisely which training regimen(s) Schieneman and Gricks advocate (or discourage) when they discuss "the manner in which the training set or seed set of documents is generated, *i.e.*, through judgmental selection *or* random selection."[34] The choice between judgmental and random selection is a false dichotomy because neither method must be used to the exclusion of the other, and because there are other methods—such as active learning—that are neither random nor judgmental. It is possible to read Schieneman and Gricks as (i) requiring exclusively random selection (and therefore prohibiting both judgmental and other non-random selection methods), or (ii) permitting judgmental selection or other non-random methods only if used in conjunction with random selection.

---

29. The most common impetus for collection enrichment (*i.e.*, culling) is to reduce the costs arising from volume-based vendor pricing, not to improve statistical estimates. *See, e.g.*, Biomet's Submission in Support of Its Discovery Efforts at 19, *In Re: Biomet M2a Magnum Implants Prods. Liab. Litig.*, No. 3:12-md-2391 (RLM) (CAN) (N.D. Ind. Apr. 4, 2013) ("Biomet has already spent over $1 million on conducting predictive coding on the 2.5+ million documents selected by search terms. Expanding predictive coding to the remaining 15.5+ million documents would cost between $2 and $3.25 million more in processing costs. . . ." (citations omitted)).

30. Schieneman & Gricks, *supra* note 1, at 249.

31. *Id.* at 251.

32. *Id.*

33. *Id.* at 249. *See also id.* ("Rule 26(g) requires counsel to exercise greater care in the collection of ESI, in order to optimize the fraction of relevant documents processed into the tool.").

34. *Id.* at 259 (emphasis added).

Regardless of which interpretation is intended, Schieneman and Gricks' thesis is predicated on the belief that non-random training regimens use training documents that may be "less than representative of the entire population of relevant documents," and therefore "run afoul of Rule 26(g)."[35] Schieneman and Gricks posit further that "[o]ne of the ways to avoid disagreement over the methodology used to train the tool" is "to share the training set with opposing counsel."[36] However, they assume, without evidence, that (i) a suitable seed[37] or training set for TAR must "reflect[ ]" the "full spectrum of relevant documents";[38] (ii) a random seed or training set is necessarily suitable (whereas a judgmentally selected set is not);[39] and (iii) it is possible for the requesting party to discern from the contents of a seed or training set whether or not it is "representative of the entire population of relevant documents," or otherwise suitable for training.[40]

The notion that seed or training sets must be random appears to derive from a false equivalence between random sampling for the purpose of *statistical estimation* and the random selection of examples for the purpose of *training a learning algorithm*.[41] It does not follow that because random sampling is necessary for statistical estimation, it is also necessary for training a learning algorithm. The argument to that effect is akin to saying that because a hammer is the proper tool to drive in a nail, it should also be used to pound in a screw.

A recent experimental study by the authors "lends no support to the proposition that seed or training sets must be random; to the contrary, keyword seeding, uncertainty sampling,[42] and, in particular, relevance feedback[43]—all non-random methods—improve significantly [at the 99% confidence level] upon random sampling."[44] Specifically, CAL—in

---

35. *Id.* at 260–61.

36. *Id.* at 261.

37. Although acknowledging ambiguity in the meaning of the phrase "seed set," Schieneman and Gricks never define the sense in which they use the phrase, and gloss over the distinction between seed set and training set. *See id.* at 259–61. As illustrated in Sections II and III, the seed set constitutes only a small fraction of the training set when using CAL, whereas the seed set is often taken to be the entire training set when using SPL. Schieneman and Gricks' conflation of seed set and training set is consistent with our impression that, in their arguments, they have considered only SPL, overlooking CAL and potentially other TAR methods.

38. *Id.* at 261.

39. *Id.* at 260–61.

40. *Id.*

41. The passage from the *TAR Glossary* defining "Judgmental Sampling" that is quoted by Schieneman and Gricks (*supra* note 1, at 260), concerns *statistical estimation*, not machine learning. *See TAR Glossary*, *supra* note 3, at 21 ("Unlike a Random Sample, the *statistical properties* of a Judgmental Sample may not be extrapolated to the entire Population." (emphasis added)).

42. "Uncertainty Sampling" is defined as "An Active Learning approach in which the Machine Learning Algorithm selects the Documents as to which it is least certain about Relevance, for Coding by the Subject Matter Expert(s), and addition to the Training Set." *TAR Glossary*, *supra* note 3, at 33-34.

43. "Relevance Feedback" is defined as "An Active Learning process in which the Documents with the highest likelihood of Relevance are coded by a human, and added to the Training Set." *Id.* at 28.

44. Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, in PROCEEDINGS OF THE 37TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION

which *none* of the training documents were randomly selected—substantially outperformed SPL—in which *all* of the training documents were randomly selected.[45] On every one of eight review tasks, the CAL approach—using keyword seeding and active learning—found more responsive documents, with less review effort, than SPL.[46] This result is partially illustrated in Table 1, which shows, for CAL and SPL, the total review effort required to achieve a recall of 75%.[47]

Table 1: Review Effort to Achieve 75% Recall

	CAL	SPL		
Matter	**Total Effort**	**Training Phase Effort**	**Review Phase Effort**	**Total Effort**
201	**6,000**	44,000	12,000	**56,000**
202	**11,000**	7,000	19,000	**26,000**
203	**6,000**	9,000	90,000	**99,000**
207	**11,000**	9,000	26,000	**35,000**
A	**11,000**	27,000	58,000	**85,000**
B	**8,000**	7,000	13,000	**20,000**
C	**4,000**	3,000	4,000	**7,000**
D	**18,000**	6,000	31,000	**37,000**

Total Effort is measured in terms of the number of documents reviewed to achieve the target recall of 75%. For CAL, Total Effort includes both training and review, which are not separate phases. For SPL, training and review are distinct phases; thus, Total Effort reflects the sum of both.

Through testing variants of CAL and SPL, the study also determined that "a simple keyword search, composed prior to the review,

(footnote continued)

RETRIEVAL (SIGIR '14) (July 2014), at 9, authors' copy *available at* http://cormack.uwaterloo.ca/cormack/calstudy/ (hereinafter "Cormack & Grossman").

45. *Id.* at 4, 5 Figure 1, 7 Table 5.

46. *Id.*

47. The Total Effort data reflected in columns 2 and 5 of Table 1 are taken from Cormack & Grossman, *supra* note 44, at 7 Table 5. The Training and Review Phase Effort data reflected in columns 3 and 4 of Table 1 are on file with the authors. For SPL, the optimal training-set size was chosen with the benefit of hindsight: The experiment was repeated with 100 training-set sizes and the result yielding the *least* total effort is reported in Table 1. In practice, the optimal training-set size would not be known in advance; it would be necessary to achieve stabilization (as described in Section III), entailing effort and uncertainty. The results reported here give SPL the benefit of the doubt, assuming that stabilization occurred at exactly the optimal training-set size, and that no extra documents were reviewed to achieve stabilization.

contributes to the effectiveness of all of the TAR protocols investigated in this study."[48]

These results call into question Schieneman and Gricks' thesis, however interpreted. For CAL, judgmental selection of the initial training set (the seed set) "generally yields superior results to random selection."[49] Active learning—a non-random method—works better than random selection for the rest of the training examples.[50] For SPL, using *some* judgmentally selected training examples improves on using entirely randomly selected ones, while using *entirely* judgmentally selected training examples often improves on using entirely randomly selected ones.[51] Overall, judgmental selection and active learning combined—without any random selection—work best of all.[52]

Our results are not entirely inconsistent with one premise behind Schieneman and Gricks' thesis: Using judgmental sampling *alone* may be a risky proposition. Indeed, the widely held belief—based on experience with some TAR tools that use SPL—that the judgmental seed set constitutes the *only* training input to the TAR tool appears to be the source of much of the angst that has been expressed regarding the use of all TAR tools, even those that also use active learning. Perhaps a gifted searcher (a "TAR Whisperer") could judgmentally select precisely the right set of examples to correctly train the SPL tool, but can the typical responding party be relied upon to conduct an adequate search? Accordingly, the argument goes, all human judgment should be eschewed in favor of random sampling. While the fear of *exclusively* judgmental selection may be valid, the proposed remedy is not.

Schieneman and Gricks' suggestion, shared by others, that "[o]ne of the ways to avoid disagreement over the methodology used to train the tool" is "to share the training set with opposing counsel,"[53] is, we believe, ill-founded. Disclosure of the seed or training set offers false comfort to the requesting party, in much the same way that disclosure of the keywords to be used for culling—without testing them—provides limited information. An examination of the responsive training documents would reveal inadequacies only to the extent that the requesting party was aware of the "full spectrum of relevant documents"[54] in the collection. But if the requesting party had such knowledge, it could more fruitfully examine the production itself, as opposed to the training set. More importantly, however, alleged "gaps" in the training set do not necessarily translate into gaps in the production, as an effective active-learning tool should readily fill those gaps.[55]

---

48. Cormack & Grossman, *supra* note 44, at 7.
49. *Id.*
50. *Id.* at 1.
51. *Id.* at 7, 8.
52. *Id.* at 4, 5 Figure 1, 7.
53. Schieneman & Gricks, *supra* note 1, at 251.
54. *Id.* at 261.
55. As evidenced by high levels of recall achieved across multiple review tasks, CAL was able to uncover documents representative of the "full spectrum of relevant documents" using a seed

Examination of the non-responsive training documents offers even less probative value than examination of the responsive training documents. Without access to the TAR tool and the collection, we are aware of no way to predict the effect of training-set selection on the effectiveness of the review. The most effective way to validate the adequacy of training-set selection—or any other choice taken in deploying TAR—is a combination of (i) prior scientific validation of the TAR tool, (ii) assurance of its proper application by qualified individuals in a given matter, and (iii) proportionate *post hoc* validation of the end-to-end review process, as discussed in Sections VII and VIII.

There are certainly steps that can be taken to afford the requesting party a measure of insight into and control over the TAR process, and therefore comfort in the judgmental selection of training examples. An obvious way is disclosure by the responding party of the search terms used to find relevant training examples, or the use of search terms specified by (or jointly developed with) the requesting party for this purpose. It is not necessary to include all of the search-term "hits" in the training set; empirical evidence shows that a random sample of the hits is sufficient to "kick-start" the learning algorithm.[56] We believe that "cherry-picking" of training examples is a questionable practice due to its unpredictable impact on the learning algorithm. Therefore, to provide adequate reassurance to the requesting party, all documents examined as potential training examples—however selected—should be coded as responsive or not, and all documents that are coded—both those that are responsive and those that are not—should be included in the training set, not just those that a TAR Whisperer deems to be "good training examples."

Empirical evidence shows that certain active-learning methods, such as CAL, are able to do a thorough job of identifying relevant documents, even when training is accomplished through non-random methods.[57] The judgmental seed set merely provides a hint to get the TAR tool started; it no more "biases" the TAR tool from finding relevant documents than driving in the direction of one's destination "biases" a GPS navigation system from finding the way. While a better seed set may improve the *efficiency* with which active learning discovers relevant documents, our research shows that, even with a seed set selected solely on the basis of a primitive keyword search, active learning is more effective and efficient than the available alternatives.[58]

(footnote continued)

set generated only from an unsophisticated keyword search. *See* Cormack & Grossman, *supra* note 44, at 7, 8.

56. *See id.* at 4.
57. *See generally id.*
58. *See generally id.*

VI. VALIDATION OF TAR OR VALIDATION OF THE END-TO-END REVIEW?

Schieneman and Gricks assert that the TAR process—which they characterize as starting once the collection has been "optimize[d] and manage[d]," and ending once a "review set" has been created[59]—must be validated by ensuring, through statistical sampling, that the review set has achieved certain precision[60] and recall targets whose values are established through a process they refer to as "stabilization."[61]

Not all TAR processes—CAL being a notable exception—involve stabilization, or the creation of a review set. Validation, we argue, should apply to the *end-to-end review*, starting with the original collection and ending with the production set, regardless of which, if any, of the steps are deemed to be "TAR." That is, validation must account for all responsive documents excluded by the review, whether before, during, or after "TAR"; or even when traditional review methods are applied.

Even if each phase, alone, excludes relatively few responsive documents, the combined effect can be substantial. Let us suppose in our hypothetical matter that the enrichment process were to discard 30% of the relevant documents (*i.e.*, were to have a recall of 70%), the TAR process were to have a recall of 75%, and the final human review were to have a recall of 70%. These numbers, in isolation, might be considered reasonable, but consider the combined effect of all three phases on our example. Of 10,000 responsive documents in the original collection, 7,000 (*i.e.*, 70%) would be retained in the enriched collection. Of the 7,000 responsive documents in the enriched collection, 5,250 (*i.e.*, 75%) would be retained in the review set. Of the 5,250 responsive documents in the review set, assuming that none were withheld as privileged, 3,675 (*i.e.*, 70%) would be identified for production. Accordingly, the recall of the end-to-end review effort would be 36.75% (*i.e.*, 3,675 of the original 10,000 responsive documents). This is a far cry from the end-to-end recall results demonstrated at TREC 2009.

We argue that the sequential application of culling methods—in this instance, enrichment, the "technology-assisted review process" (as defined by Schieneman and Gricks[62]), and the subsequent manual review of the review set—will, due to the multiplier effect described above, generally yield inferior recall. Prior enrichment is unnecessary, as discussed in Section IV, and the subsequent human review effort will

---

59. Schieneman & Gricks, *supra* note 1, at 251, 269.

60. We mention precision only in passing because, as noted by Schieneman and Gricks, precision is less informative than recall, and is not likely to be an issue in most reviews. *Id.* at 268 ("As a practical matter, the precision of technology-assisted review processes is much greater than the precision of other review alternatives, making it unlikely that irrelevant documents are being produced for some improper purpose in violation of Rule 26(g). Moreover, the documents identified for production through the technology-assisted review process are generally reviewed before production, further minimizing the production of irrelevant documents in violation of Rule 26(g)." (citations omitted)).

61. *Id.* at 263.

62. *Id.* at 269.

introduce its own error.[63] Whether or not several culling phases are used, validation should estimate how many of the responsive documents in the *original collection* are being *produced*, not how many of the responsive documents in the *enriched collection* are being *reviewed*. Extraordinary efforts to achieve the latter, we argue, often come at the expense of the former.

VII. VALIDATION METHODS

Schieneman and Gricks narrowly equate validation with *post hoc* statistical sampling to estimate recall, excluding other forms of evidence pertinent to the efficacy of a review. Furthermore, they assert, without support, that certain sampling methods provide accurate and efficient recall estimates.

We introduce, in Subsection A, the framework for our argument that prior evidence of a method's effectiveness, coupled with evidence that the method was properly applied by qualified individuals, is at least as important as *post hoc* validation. We argue, in Subsection B, that recall, while an intuitively appealing and reasonably informative measure of review effectiveness, does not tell the whole story, and is difficult to measure accurately or consistently. We show, in Subsection C, that *post hoc* sampling methods for recall estimation in common use today—including those advanced by Schieneman and Gricks, and others—are either mathematically unsound or could require the manual review of unreasonably large samples, resulting in disproportionate effort.

A. Prior and *Post Hoc* Evidence

Validating a review effort involves consideration of all available evidence, including prior scientific evidence confirming the validity of the method, evidence that the method was properly applied by qualified individuals, and subsequent evidence that readily observable phenomena are consistent with a successful outcome.

When cooking a turkey, one can be reasonably certain that it is done, and hence free from salmonella, when it reaches a temperature of at least 165 degrees throughout. One can be reasonably sure it has reached a temperature of at least 165 degrees throughout by cooking it for a specific

---

63. *See generally, e.g.*, Maura R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267 (2012), *available at* http://digitalcommons.pace.edu/plr/vol32/iss2/1/; William Webber, Douglas W. Oard, Falk Scholer & Bruce Hedin, *Assessor Error in Stratified Evaluation*, in PROCEEDINGS OF THE 19TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM '10) 539 (Oct. 2010), *available at* http://www.williamwebber.com/research/papers/wosh10_cikm.pdf; Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70 (2010); Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697 (2000), *available at* http://dx.doi.org/10.1016/S0306-4573(00)00010-8 (hereinafter "Voorhees").

amount of time, depending on the oven temperature, the weight of the turkey, and whether the turkey is initially frozen, refrigerated, or at room temperature. Alternatively, when one believes that the turkey is ready for consumption, one may probe the turkey with a thermometer at various places. Both of these approaches have been validated by biological, medical, and epidemiological evidence. Cooking a turkey requires adherence, by a competent cook, to a recipe that is known to work, while observing that tools like the oven, timer, and thermometer appear to behave properly, and that the appearance, aroma, and texture of the turkey turn out as expected. The totality of the evidence—vetting the method in advance, competently and diligently applying the method, and monitoring observable phenomena following the application of the method—supports the reasonable conclusion that dinner is ready.

B. Recall Uncertainties

Recall can be an informative indicator of the completeness of a review effort, but it is difficult to measure properly and can be misleading. Scientific studies like TREC have expended vast resources—far more than would be reasonable and proportionate in most matters—measuring the recall of various information-retrieval approaches. Even so, the margins of error at the 95% confidence level, used at TREC 2009, and considered in the JOLT Study, ranged from ±5.7% to ±25.9%[64]—larger than the maximum proposed by Schieneman and Gricks. For many matters, it is neither feasible nor necessary to expend the effort to estimate recall with a margin of error of ±5%, at a 95% confidence level—the standard required for scientific publication.

Margins of error and confidence levels quantify only one source of uncertainty—random error, or uncertainty due to chance—in estimating recall and other statistics. A recall estimate, however, is meaningless without an unambiguous characterization of (i) the set of ESI over which the recall is calculated, and (ii) responsiveness. As noted in Section IV, collection enrichment reduces the set of documents subject to review, potentially yielding a higher and easier-to-calculate recall estimate that is not comparable to the recall estimate that would be obtained based on the original collection. At its core, collection enrichment to facilitate validation is itself an exercise in judgmental sampling from which, as Schieneman and Gricks acknowledge,[65] no valid statistical inference can be drawn.

---

64. Hedin et al., *supra* note 21, at 17 Table 6 (Topics 201, 202, 203, and 207). The margins of error may be derived from the confidence intervals ("95% C.I.") shown in Table 6. For a method of computing somewhat smaller margins of error, *see* William Webber, *Approximate Recall Confidence Intervals*, 31 ACM TRANSACTIONS ON INFO. SYS. 1 (Jan. 2013), preprint *available at* http://arxiv.org/pdf/1202.2880v1.pdf (hereinafter "Webber").

65. Schieneman & Gricks, *supra* note 1, at 260 ("A method in which a sample of the document population is drawn, based at least in part on subjective factors, so as to include the 'most interesting' documents by some criterion; the sample resulting from such a method. *Unlike a*

To define any set, it is necessary to define what constitutes an element of the set. In the case of ESI, this choice can be surprisingly complicated and controversial. Are duplicate documents distinct elements of the set, or the same element? Are email messages and their attachments separate elements, or is a message with all of its attachments, one element? Even the TREC 2009 coordinators were unable to resolve the latter issue and reported two sets of substantially different recall estimates:

1. Message-level recall, in which each email message, with attachments, was considered to be a single element. For the efforts reported in the JOLT Study, message-level recall ranged from 67.3% to 86.5%, with an average of 76.8%;[66]

2. Document-level recall, in which each email and its attachments were considered to be distinct elements. For the efforts reported in the JOLT Study, document-level recall ranged from 76.1% to 89.6%, with an average of 84.1%.[67]

Unfortunately, much commentary, many pleadings, and some opinions cite aspirational or measured recall values—70% or 75%—typically citing the TREC 2009 *message-level* results for support, without specifying the set of ESI over which their own recall is calculated (often what appears to be *document-level* recall). This is an invalid comparison.

Defining responsiveness is equally problematic. While requests for production ("RFPs") are typically specified in writing, they are subject to interpretation, and reasonable, informed reviewers may disagree on their interpretation, or on how to apply that interpretation to determine whether or not any particular document is responsive. Even a single reviewer may make a different determination, under different circumstances, including his or her knowledge at the time of the determination, arguments or incentives to decide one way or the other, and human factors such as fatigue or distraction, which can result in clerical errors. Experts disagree surprisingly often with each other, and even with themselves.[68]

Uncertainty in determining responsiveness necessarily limits the ability to measure recall. This limitation can be profound. At TREC 2009, "pre-adjudication scores" (*i.e.*, metrics computed prior to

(footnote continued)

random sample, the statistical properties of a judgmental sample may not be extrapolated to the entire population." (emphasis in original) (quoting *TAR Glossary*, *supra* note 3, at 21)).

66. Hedin et al., *supra* note 21, at 17 Table 6 (Topics 201, 202, 203, and 207).

67. National Institute of Standards and Technology, *Per-Topic Scores: TREC 2009 Legal Track, Interactive Task*, in NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS, Appendix titled *Legal Interactive Run Results*, at 3, 4, 5, 6, 9 Tables 4, 8, 12, 16, 28 (Document-based, Post-Adjudication Scores for Topics 201, 202, 203, 204, and 207), *available at* http://trec.nist.gov/pubs/trec18/appendices/app09int2.pdf.

68. *See generally supra* note 63.

undertaking the quality-assurance process), including recall, were determined using the coding of third-party reviewers (either law students or contract reviewers). The pre-adjudication message-level recall for the efforts reported in the JOLT Study ranged from 3.6% to 70.9%, with an average of 24.3%[69]—more than 50% lower than the official recall scores reported above after quality-assurance efforts. The official recall scores were achieved through the use of an extraordinary appeal and adjudication process used to correct coding errors by the third-party reviewers. The bottom line is that *inconsistencies in responsiveness determinations limit the ability to estimate recall*. Even a perfect review, performed by an expert—with a recall of 100%—is unlikely to achieve a measured recall of higher than 70%, if the final responsiveness assessment is made by a second, independent expert.[70] If the responsiveness assessment is made by an inexpert reviewer, chances are that the measured recall will be considerably lower, as evidenced by the TREC 2009 "pre-adjudication scores."

There is not necessarily a one-to-one correspondence between the definition of responsiveness and a single written RFP. Typically, a discovery request or subpoena will include at least several, and sometimes dozens, of RFPs. Often, a number of RFPs will seek similar subject matter from the same universe of ESI, and it would therefore be expedient to use a single, combined review (or a small number of combined reviews) to find documents responsive to all of these RFPs. As far as the review is concerned, a document is responsive if it is responsive to at least one of the RFPs. It is not obvious that a single recall estimate based on this amalgamated definition of responsiveness is a good indicator of the thoroughness of the review. Let us suppose that, in our hypothetical matter, the discovery request were to contain three RFPs ("RFP 1," "RFP 2," and "RFP 3"). Suppose further that, of the 10,000 responsive documents, 9,000 were responsive to RFP 1, 900 were responsive to RFP 2, and 100 were responsive to RFP 3. A reported recall of 70% might indicate that 6,300 documents responsive to RFP 1, 630 documents responsive to RFP 2, and 70 documents responsive to RFP 3, were found (*i.e.*, there is 70% recall on each RFP). A reported recall of 70% might equally well indicate that, by accident or design, 7,000 documents responsive to RFP 1, and none responsive to the other two RFPs, were found. In the latter case, "70% recall" presents as reasonable a clearly unreasonable result. We argue that, in order to be

69. National Institute of Standards and Technology, *supra* note 67, at 3, 4, 5, 6, 9 Tables 1, 5, 9, 13, 25 (Message-based, Pre-Adjudication Scores for Topics 201, 202, 203, 204, and 207).

70. Maura R. Grossman, Gordon V. Cormack, Bruce Hedin & Douglas W. Oard, *Overview of the TREC 2011 Legal Track*, in NIST SPECIAL PUBLICATION: SP 500-296, THE TWENTIETH TEXT RETRIEVAL CONFERENCE (TREC 2011) PROCEEDINGS, at 9 (2011), *available at* http://trec.nist.gov/pubs/trec20/papers/LEGAL.OVERVIEW.2011.pdf ("Somewhat higher recall may be achieved with more effort, but it is unclear whether improvements in recall measures above 70% are meaningful, given the inherent uncertainties arising from sampling and human assessment of responsiveness."); Voorhees, *supra* note 63, at 701 ("The scores for the two sets of secondary judgments imply a practical upper bound on retrieval system performance of 65% precision at 65% recall since that is the level at which humans agree with one another.").

considered reasonable, the review must be shown to find material responsive to *each* of the RFPs (assuming the collection contains documents responsive to each), which a single recall measure cannot do. On the other hand, separately estimating the recall for each individual RFP is likely to require disproportionate effort. Statistics cannot solve this problem, and therefore other approaches may be warranted.

By similar reasoning, we argue that, to be considered reasonable, a review should be shown to cover all aspects of responsiveness, even if each aspect is not explicitly specified in a separate RFP. For example, if the collection were to include email messages, PDFs, word-processing documents, and spreadsheets containing potentially responsive information, no one would question that the review should cover all of these file types. By the same token, if the RFP were to request documents reflecting "actions by any board member," the review should be performed to cover all board members. In general, a single recall estimate may mask the fact that some identifiable sub-population of documents has been searched inadequately, or missed entirely.

Recall further masks the issue that responsive documents differ in their importance in resolving the issues in dispute. If the information contained in the 30% of the responsive documents omitted by a review achieving 70% recall is largely duplicative or unimportant, the review is qualitatively superior to another with the same recall that omits a large amount of non-duplicative, important information.

Regardless of the recall estimate, it behooves the responding party to investigate whether the documents that are missed form an identifiable sub-population, or are non-duplicative and important. If so, the review should continue. On the other hand, if a review strategy has been shown in advance to be effective, is properly applied by qualified individuals to the matter at hand, and observable indicators are consistent with having adequately searched for all aspects of responsiveness and all identifiable sub-populations within the collection, then a less precise or less burdensome recall estimate that yields a result consistent with a reasonable search should be sufficient. Extraordinary efforts to estimate recall—at the standard required for scientific publication—are unwarranted and disproportionate in light of the totality of the evidence.

C. *Post Hoc* Recall Estimation Methods

Uncertainty in recall estimates arises from both random error (*i.e.*, chance), and also from inconsistency in relevance determinations. An estimate should minimize the overall uncertainty from both sources, for a given level of estimation effort. In the proportionality equation, effort to reduce uncertainty in estimation competes for resources with effort to improve the review process. It is therefore important to use an efficient estimation method, and to expend resources to measure recall only as necessary to satisfy the obligation of a reasonable inquiry under Rule 26(g). Indeed, there may be some situations where it is disproportionate

to compute a statistical recall estimate at all, in light of the cost of sampling and the availability of other methods of validation.

A common way to estimate recall is by *post hoc* analysis, in much the same way that a thermometer is used to determine that a turkey is cooked before it is served. A number of TAR protocols apply statistical analyses repeatedly to track the progress of the review,[71] in much the same way that a thermometer can be used over and over to track the progress of cooking a turkey. Whether employed *post hoc*, or in an ongoing fashion to track progress, many of the estimation methods proposed or commonly in use today are either statistically unsound or, in our opinion, disproportionately burdensome.

Schieneman and Gricks discuss four statistical methods for computing *post hoc* recall estimates:[72] the "Direct Method," proposed by David D. Lewis on behalf of plaintiffs in the *Kleen Products* case,[73] and ordered by consent in the *In re Actos* case,[74] and three methods that estimate recall as the ratio of two separate statistical estimates (together, the "Ratio Methods"). The Direct Method, according to Schieneman and Gricks, may involve disproportionate effort, in light of their assertion that "there are alternative means of calculating recall that do not require such a significant effort"[75] (*i.e.*, the Ratio Methods). While we agree that employing the Direct Method will often entail disproportionate effort, we also argue that the purported advantages of the Ratio Methods are largely illusory, predicated as they are on unsound mathematics.

The Direct Method of estimating recall proceeds as follows. Documents are drawn repeatedly, at random, from the collection, and are reviewed for responsiveness until 385 responsive documents are identified. After—and only after—the 385 responsive documents have been identified, it is determined what percentage of the 385 documents had previously been identified as responsive by the TAR process, however defined. This percentage is the recall estimate, with a margin of error of ±5%, and 95% confidence. The Direct Method is statistically

---

71. Repeated statistical estimates may involve drawing repeated samples (as discussed in Section III), or more commonly, drawing a sample at the outset (a "control set"), and repeating the statistical calculations using the control set as a sample. Problems with the use of repeated estimates to track progress, and to determine when recall is "adequate," are addressed elsewhere. *See, e.g.*, Mossaab Bagdouri, William Webber, David D. Lewis & Douglas W. Oard, *Toward Minimizing the Annotation Cost of Certified Text Classification*, in PROCEEDINGS OF THE 22ND ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT (CIKM '13) 989 (Oct. 2013), *available at* http://dl.acm.org/citation.cfm?doid=2505515.2505708; William Webber, Mossaab Bagdouri, David D. Lewis & Douglas W. Oard, *Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness*, in PROCEEDINGS OF THE 36TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR '13) 933 (July 2013), authors' version *available at* http://terpconnect.umd.edu/~oard/pdf/sigir13webber.pdf.

72. Schieneman & Gricks, *supra* note 1, at 272–73.

73. Hr'g Tr. at 259–64, *Kleen Prods., LLC* v. *Packaging Corp. of Am.*, No. 10-C-5711 (N.D. Ill. Feb. 21, 2002).

74. Case Mgmt. Order: Protocol Relating to the Produc. of ESI, *In Re: Actos (Pioglitazone) Prods. Liab. Litig.*, MDL No. 6:11-md-2299 (W.D. La. July 27, 2012).

75. Schieneman & Gricks, *supra* note 1, at 273.

sound, but is quite burdensome, especially when richness is low. In our hypothetical matter, it would be necessary to draw and manually review about 38,500 documents to calculate the recall estimate—nearly twice as much effort as would be involved in the entire end-to-end CAL process.[76] Were the richness lower—as it was for four of the review tasks considered in the JOLT Study—the burden would be even more onerous. In general, the Direct Method entails the manual review of 385/R documents, where R is the richness of the collection. For each of the two JOLT Study tasks for which richness was 0.3%, the Direct Method would require the review of 385/0.003=128,333 documents to compute the recall estimate. The evidence that would be educed from such an estimate is clearly not worth the candle.

Moreover, if the Direct Method were employed in our hypothetical matter, and were to yield a recall estimate of 75%, it would follow that 25% of the 385 documents—96 in total—would be responsive, but not identified by the TAR process. These documents would obviously need to be produced, but also examined to determine whether they contained non-duplicative, important information. The documents would likely represent a viable seed or training set for a supplemental TAR search. In contrast to sequential culling, the cumulative effect of sequential searching is beneficial. If the supplemental search—applied only to the documents not yet identified as responsive—had a recall of just 60%, it would find 1,500 of the 2,500 remaining responsive documents, for a total of 9,000 documents, or 90% overall recall.

Schieneman and Gricks assert that two Ratio Methods, a basic method (the "Basic Ratio Method") and the method employed in the *Global Aerospace* case[77] (the "*Global* Method"), achieve an acceptable recall estimate, with less sampling effort, than the Direct Method. Although dismissed by Schieneman and Gricks as "unnecessarily indirect,"[78] a third Ratio Method ("eRecall") has been advanced by Herbert L. Roitblat to address the same proportionality concerns.[79] The foundation for these assertions appears to be the following argument:

---

76. *See supra* p. 291 and note 26.

77. Letter from Gordon S. Woodward, Att'y for Landow Entities, to All Counsel, *Global Aerospace Inc.* v. *Landow Aviation, L.P.*, Consol. Case No. CL601040 (Va. Cir. Ct. Nov. 30, 2012) ("'esi' Sampling Report" on docket).

78. Schieneman & Gricks, *supra* note 1, at 272.

79. *Measurement in eDiscovery*, *supra* note 15, at 7–10 ("In order to find 400 responsive documents at 10% Richness or Prevalence, we would have to sample approximately 4,000 documents, randomly chosen from the collection as a whole, without regard to whether they were predicted to be responsive or not (10% of 4,000 is 400). That's a lot of work, and it may be more than the number of documents needed to train the process in the first place (if we are using predictive coding).

If Prevalence is lower, if only a small percentage of documents is actually responsive, measuring Recall directly can be even more costly, because a still large sample of random documents would have to be examined. Calculating Recall directly can be done, but it takes a very substantial amount of work just to find the responsive documents to measure. . . .

Fortunately, there are other ways to assess whether we conducted a reasonable inquiry with far less effort.").

> *The Ratio Method Fallacy*: A sample of size 385 yields a margin of error of ±5%, with 95% confidence. Therefore, if we use a sample of size 385 to estimate the number of responsive documents in the *production*, and another sample of size 385 to estimate the number of responsive documents in the *collection*, the ratio of these two estimates is a valid recall estimate with a margin of error of ±5%, and a confidence level of 95%.

This argument is specious. The Ratio Methods differ from one another only in how they estimate the number of responsive documents in the production, and the number of responsive documents in the collection, prior to dividing the former by the latter to yield an estimate of recall. The Basic Ratio Method samples the *production* and the *collection*;[80] the *Global* Method samples the *production* and the *null set*;[81] while eRecall samples the *null set* and the *collection*.[82]

In each case, the sample estimates are combined using simple algebra to form an estimate of recall. However, in each case, the algebra yields a biased[83] point estimate, with a margin of error dramatically larger than ±5% and/or a confidence level dramatically lower than 95%.

While the correct calculations are complex,[84] it is a straightforward matter to demonstrate that the calculations underlying *The Ratio Method Fallacy* are incorrect. From the definitions of point estimate,[85] margin of error,[86] and confidence level,[87] we know that if the sampling is repeated a large number of times, the average of an unbiased point estimate of recall should approach the true value, and the point estimate should be within ±5% of the true value, 95% of the time. To validate each estimation method, we simulated its application 100,000 times to our hypothetical production set known to have a recall of 75%. If valid, the estimation

---

80. Schieneman and Gricks, *supra* note 1, at 273.

81. *Id.* The "Null Set" is defined as "The set of Documents that are not returned by a search process, or that are identified as Not Relevant by a review process." *TAR Glossary*, *supra* note 3, at 25.

82. *Measurement in eDiscovery*, *supra* note 15, at 7–10. *See also* Herbert L. Roitblat, *A Tutorial on Sampling in Predictive Coding* (OrcaTec LLC 2013), at 3, *available at* http://orcatec.com/2013/10/07/a-tutorial-on-sampling-in-predictive-coding/.

83. A statistic is biased if the long-term average of the statistic is not equal to the parameter it is estimating. A biased estimation method, when repeated a large number of times, will yield an estimate that is, on average, higher or lower than the true value. In contrast, an unbiased estimation method, if repeated a large number of times, will, on average, be indistinguishable from the true value. *See* MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 3 (2d ed. 2001) ("An estimate[e] is said to be unbiased if its average over all possible random samples equals the population parameter no matter what that value may be.").

84. *See generally* Webber, *supra* note 64.

85. "Point Estimate" is defined as "The most likely value for a Population characteristic. When combined with a Margin of Error (or Confidence Interval) and a Confidence Level, it reflects a Statistical Estimate." *TAR Glossary*, *supra* note 3, at 25.

86. "Margin of Error" is defined as "The maximum amount by which a Point Estimate might likely deviate from the true value, typically expressed as 'plus or minus' a percentage, with a particular Confidence Level." *Id*. at 22.

87. "Confidence Level" is defined as "[T]he chance that a Confidence Interval derived from a Random Sample will include the true value." *Id.* at 12.

method should yield point estimates of recall whose average is nearly indistinguishable from 75%,[88] of which 95% fall between 70% and 80% (*i.e.*, within the margin of error of ±5%). Moreover, if the point estimate were unbiased, and the margin of error were ±5% as claimed, the 2.5th percentile of the estimates should be 70%, while the 97.5th percentile should be 80%. In other words, 2.5% of the time, the estimate should fall below 70%, and 2.5% of the time it should fall above 80%.

As seen in Tables 2 and 3, the Direct Method yields an unbiased estimate, with more than 95% of the estimates falling within the predicted margin of error. The 2.5th and 97.5th percentiles are close to the predicted values. In short, the Direct Method is statistically sound, yielding an unbiased point estimate and a margin of error slightly better than claimed.

The Ratio Methods, on the other hand, yield biased estimates with margins of error vastly larger than claimed. The Basic Ratio Method and the *Global* Method overestimate—while eRecall underestimates—the true recall value, especially, but not exclusively, when prevalence is low. None of the Ratio Methods provides recall estimates that fall within the required margin of error anywhere near 95% of the time. None of the Ratio Methods has 2.5th and 97.5th percentiles anywhere near 70% and 80%, respectively. In short, the appeal of less sampling effort[89] is belied by the fact that the estimates are unsound. On the other hand, the Direct Method, while mathematically sound, entails *much* larger samples, especially for low prevalence. In short, there is no free lunch.

Table 2: Recall Estimation Validity Results—1% Collection Richness

Method	Sample Size Required for Computation of Recall	Average (Should Be 75%)	% Within ±5% (Should Be 95%)	2.5th Percentile (Should Be 70%)	97.5th Percentile (Should Be 80%)
Direct	38,500	75.0%	97.5%	70.6%	79.2%
Basic Ratio	770	83.4%	17.4%	35.7%	100.0%
Global	770	79.4%	36.7%	48.5%	100.0%
eRecall	770	69.3%	8.9%	0.0%	100.0%
Simulated recall estimates for a hypothetical review effort with a known recall of 75%, a known precision of 83.3%, and a known prevalence of 1%.					

88. STEVEN K. THOMPSON, SAMPLING, 34 (3d ed. 2012) ("One can assess the bias or expected error by looking at the difference between the average value of the estimator over all the samples and the true test population characteristic.").

89. Schieneman & Gricks, *supra* note 1, at 273.

Table 3: Recall Estimation Validity Results—10% Collection Richness

Method	Sample Size Required for Computation of Recall	Average (Should Be 75%)	% Within ±5% (Should Be 95%)	2.5th Percentile (Should Be 70%)	97.5th Percentile (Should Be 80%)
Direct	3,850	75.0%	97.5%	70.6%	79.2%
Basic Ratio	770	76.5%	33.7%	56.8%	100.0%
Global	770	75.4%	61.5%	64.6%	86.6%
eRecall	770	74.4%	43.4%	54.7%	88.9%
Simulated recall estimates for a hypothetical review effort with a known recall of 75%, a known precision of 83.3%, and a known prevalence of 10%.					

The above examples serve to illustrate that there is no shortcut to achieving a scientific-publication-quality estimate of recall, a standard that may not be necessary or appropriate for the typical TAR review. The bottom line is that better recall estimates require more work; work that may be disproportionate to the contribution of the estimate for validation purposes. The Ratio Methods—the *Global* Method, in particular—can be adjusted to calculate an unbiased point estimate, with an appropriate confidence interval,[90] at the expense of requiring sample sizes comparable to those of the Direct Method.

It bears repeating that, regardless of sample size or sampling regimen, inconsistency in human assessment of responsiveness can result in substantial error in recall estimates, as illustrated in Subsection B. Sample-based estimation depends on human review of the sample, and even an expert reviewer will be fallible, resulting in estimation error.[91] With perhaps disproportionate effort, this source of estimation error can be reduced, but not eliminated; for example, by using a committee of experts to review the sample. With additional effort, statistical estimation error may likewise be reduced by increasing sample size indefinitely. The bottom line is that efforts to reduce one kind of error come at the expense of efforts to reduce the other, and, in any event, validation efforts compete for resources that might otherwise be used to conduct a better review.

VIII. A WAY FORWARD: COMBINING PRIOR AND *POST HOC* EVALUATION

We are unconvinced that extraordinary efforts to conduct *post hoc* recall estimates are justified for the purpose of validating a single review, for the same reason that it is not necessary to use a laboratory-quality

90. *See generally* Webber, *supra* note 64.
91. *See supra* note 70. *See also generally supra* note 63.

thermometer—or indeed any thermometer at all—to ensure with reasonable certainty that a turkey is cooked. We believe that effort spent in advance to scientifically validate the TAR method, ongoing oversight to ensure the method is properly applied by qualified individuals, and proportionate *post hoc* testing, would provide more confidence in the review effort than Herculean *post hoc* sampling efforts on a per-matter basis. In so arguing, we consider the factors summarized in the syllabus of the *Daubert* case:[92]

> Faced with a proffer of expert scientific testimony under Rule 702, the trial judge, pursuant to Rule 104(a), must make a preliminary assessment of whether the testimony's underlying reasoning or methodology is scientifically valid and properly can be applied to the facts at issue. Many considerations will bear on the inquiry, including whether the theory or technique in question can be (and has been) tested, whether it has been subjected to peer review and publication, its known or potential error rate, and the existence and maintenance of standards controlling its operation, and whether it has attracted widespread acceptance within a relevant scientific community. The inquiry is a flexible one, and its focus must be solely on principles and methodology, not on the conclusions that they generate.

To be clear, we do not suggest that a formal *Daubert* hearing is necessary or appropriate in evaluating individual TAR efforts; the admissibility of expert testimony at trial and the adequacy of production present different objectives and standards.[93] Nevertheless, both share the common concern of assessing whether a method of inquiry is reasonable, valid, and properly applied. Accordingly, even if not dispositive, the *Daubert* factors may be instructive in the TAR context. Notably, the *Daubert* test focuses on *a priori* validation—establishing, through accepted scientific standards, the applicability of the method to the task at hand, its potential error rate, and standards controlling its operation. Therefore, TAR tools and methods should be pre-established as valid in their own right, not re-established *de novo* for each and every matter. *Daubert* further emphasizes flexibility, which we interpret to mean that *all available evidence* should be considered in validating a TAR method

---

92. *Daubert* v. *Merrell Dow Pharm., Inc.*, 509 U.S. 579, 580 (1993).

93. *See Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, 189 (S.D.N.Y. 2012) ("[Federal Rule of Evidence] 702 and *Daubert* simply are not applicable to how documents are searched for and found in discovery."); *but see* David J. Waxse & Brenda Yoakum-Kriz, *Experts on Computer-Assisted Review: Why Federal Rule of Evidence 702 Should Apply to Their Use*, 52 Washburn L.J. 207, 223 (Spring 2013), *available at* http://contentdm.washburnlaw.edu/cdm/ref/collection/wlj/id/6195 ("Rule 702 and the *Daubert* standard should be applied to experts with technical expertise or knowledge pertinent to a party's ESI search and review methodologies and who provide the court with evidence on discovery disputes involving these methods.").

and its application to any particular review—not just a single bright-line test such as recall.

Notwithstanding the caveats expressed above with regard to recall—or any statistical measure—as the sole gauge of success, recall can be estimated prior to any particular review effort, based on the results of previous efforts. For example, the CAL method examined in the JOLT Study achieved document-level recall point estimates of 84.3%, 84.4%, 86.0%, and 89.6% on the four TREC 2009 review tasks to which it was applied.[94] Applying the t-distribution[95] to these estimates yields a 95% confidence interval of 82.1% to 90.0%, meaning that, if the same method were applied with equal diligence to a fifth review task, we would expect with 95% confidence to achieve, on average, a true recall between 82.1% and 90.0%. In other words, *a priori* statistical evaluation of the method on four tasks can yield as high a confidence level in estimating the result of a fifth task, as can *post hoc* evaluation of the fifth task alone. In addition, a recall point estimate computed at the end of a review—even one with a high margin of error—augments confidence, not only in the particular result, but in future results. Once a sufficient *a priori* confidence level has been established, it should be sufficient to ensure that the method is properly applied by qualified individuals, and that readily observable evidence—both statistical and non-statistical—is consistent with the proper functioning of the method.

We do not claim that the use of CAL, prior scientific validation of the TAR method, competence and oversight in its application, and proportionate *post hoc* sampling are the only ways forward, but we are unaware of any approaches that work better, either for TAR or for validation. There is much more to be learned about the application and validation of TAR, and, we argue, the legal industry would be better served by researching, developing, and implementing improvements in TAR, than by expending heroic efforts—under the guise of compliance with Rule 26(g)—to validate individual TAR efforts through the application of statistical methods that are either mathematically unsound, or so disproportionate that they create disincentives to use TAR, and threaten to eradicate the savings afforded by its use.

IX. CONCLUSION

Any effort to codify what is reasonable—under Rule 26(g) or otherwise—is necessarily limited by the difficulty of anticipating the unintended consequences of such prescriptions. We have illustrated a number of circumstances in which an obligation to follow the practices dictated by Schieneman and Gricks would preclude the use of perfectly reasonable—indeed superior—TAR methods and validation strategies.

---

94. *See supra* note 67. *See also* Cormack & Mojdeh, *supra* note 10, at 8 Table 5.

95. STEFAN BÜTTCHER, CHARLES L. A. CLARKE & GORDON V. CORMACK, INFORMATION RETRIEVAL: IMPLEMENTING AND EVALUATING SEARCH ENGINES 423 (2010).

An obligation to follow Schieneman and Gricks' prescriptions would entrench one particular approach to TAR, and stifle the state of the art.

There is no reason that the criteria for reasonableness under Rule 26(g) should be "unique" to TAR: All review methods leave responsive documents behind, and all review methods may be subject to validation—statistical or otherwise. Application of standards of reasonableness piecemeal to discrete phases of a review effort—*e.g.*, collection, disclosure, training, stabilization, and validation—has the potential to invite discovery about discovery and successive motions practice, thereby impeding the "just, speedy, and inexpensive determination of every action and proceeding."[96]

Accordingly, we believe that the Rule 26(g) requirement to conduct a reasonable inquiry may be fulfilled without undertaking the burdensome and potentially counterproductive enrichment, training, stabilization, and validation techniques asserted by Schieneman and Gricks to be obligations under Rule 26(g). Exceptional efforts to enrich the corpus have not been shown to enhance either review quality or confidence in the result. Random selection of training documents, to the exclusion of non-random methods, has not been shown to enhance either review quality or confidence in the result; to the contrary, Continuous Active Learning—a method independently shown to achieve superior results—uses a judgmental seed set and active learning, both non-random methods, and is entirely consistent with the obligations mandated by Rule 26(g). Finally, heroic efforts to measure recall for any particular review may be unduly burdensome and divert resources from conducting a better review. Such metrics can be gamed or misinterpreted, and will often require statistical sophistication beyond what might reasonably be expected of the average responding party. Most dangerous of all, however, is the risk that efforts aimed at maximizing recall estimates may drive TAR workflow choices in a direction that is inconsistent with best practice.

96. Fed. R. Civ. P. 1.